AD_____

Award Number: W81XWH-07-1-0447

TITLE: Mathematical Modeling and Analysis of Mass Spectrometry Data in Workflows for the Discovery of Biomarkets in Breast Cancer

PRINCIPAL INVESTIGATOR: Vladimir Fokin, Ph.D.

CONTRACTING ORGANIZATION: Indiana University
Indianapolis, IN 46202-5167

REPORT DATE: July 2008

TYPE OF REPORT: Annual

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
1 Jul 2008	Annual	1 Jul 2007 – 30 Jun 2008

4. TITLE AND SUBTITLE

Mathematical Modeling and Analysis of Mass Spectrometry Data in Workflows for the Discovery of Biomarkets in Breast Cancer

5a. CONTRACT NUMBER

5b. GRANT NUMBER
W81XWH-07-1-0447

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Vladimir Fokin, Ph.D.

E-Mail: vvf@math.iupui.edu

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Indiana University
Indianapolis, IN 46202-5167

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

The major achievement for the last year was the development of the workflow for the LC-MS/MS proteomic experiment and testing it on the real data. This included setting up an appropriate mathematical/statistical scheme of random-effect models. The overall significance of multiple hypothesis testing was controlled by the false discovery rate approach. To improve the statistical power and clinical interpretation of the results, the protein identifications were mapped to high quality publicly available data sources, and differential quantification was rolled up to the gene level. The technique similar to the gene set enrichment analysis in microarrays is used to find statistically significant differences between the protein families. To achieve a dataset with minimal false identifications, functional annotations from the GeneOntology and the HUPO project were added as well as known plasma concentration from the literature. To understand the biological relevance of the differentially expressed proteins we used several functional annotation tools. We plan to complement the mass-spectrometry data with metabolomic analysis, and will provide additional data from protein microarrays.

15. SUBJECT TERMS
Mass-spectrometry proteomics, plasma proteomics, LC-MS/MS, cancer proteins.

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			19b. TELEPHONE NUMBER *(include area code)*
U	U	U	UU	13	USAMRMC

Table of Contents

Introduction:

Develop the mathematical tools into the workflow for the analysis of plasma/serum samples obtained from the large clinical trials, incorporate the additional biological information from different sources. Incorporate this developed techniques into the Data Integrative Specimen Information Services including visualization tools to aid clinicians in the decision making.

Body:

Multidisciplinary Training Program

During the first year of the grant (and while waiting for financing to begin), I have finished most of the required formal training planned for the whole duration of the grant, thus bringing the total number of post-graduate courses and workshops to over 15 in the areas of clinical research, molecular biology, statistics, and computer sciences. Since my first submission for this grant I have successfully completed the following courses and workshops (according to the **Training Tasks, TT** of the Statement of Work):

GRAD-G 890 Methods in Molecular Biology and Pathology, **TT1**
STAT 598 (STAT 521) Modern Statistical Computing, **TT2**
GRAD-N 802 Techniques of Effective Grant Writing, **TT4**
GRAD-G 653 Intro to Applied Multivariate Statistical Methods, **TT2**
STAT 514 Design of Experiment, **TT3**
STAT 523 Categorical Data Analysis, **TT3**
STAT 528 Intro to Mathematical Statistics, **TT2**
STAT 598 (BIOS 627) Statistical Issues in Pharmaceutical Research, **TT2**
Proteomics Informatics course, Institute for System Biology, Seattle, WA, May 6-11, 2007, **TT1**
Bioconductor Advanced Course, Northwestern University, Chicago, IL, October 1-3, 2007, **TT3**
Introduction to R and Bioconductor, Fred Hutchinson Cancer Research Center, Seattle, WA, November 28-30, 2007, **TT3**
MS/MS: Introduction, Short course at American Society of Mass Spectrometry Annual Conference, Denver, CO, May 31, 2008 – June 1, 2008, **TT1**

I have been closely working with Susanne Ragg, M.D., Ph.D. of Indiana University School of Medicine, and Gunther Shadow, M.D., Ph.D. of Regenstrief Institute and Indiana University School of Informatics who performed the direct mentoring and reporting to Dr. McDonald and Dr. Sledge. Dr. Ragg is also director of Center for Computational Diagnostics, and it provided me with close collaborations with researches from different areas of science and medicine. We have met weekly at the Center for Computational Diagnostics meetings to discuss the progress. We have prepared several presentations at distinct meetings, including my poster presentation at the "Era of Hope" meeting in Baltimore, MD. My mentor in statistics, Krzysztof Podgorski, Ph.D., has moved to accept the chair position of the Department of Mathematics and Statistics at Lund University, Sweden, while continuing the mentoring and guidance for me through e-mail, phone, and Skype communications. As the result of this work we are preparing a publication on the alignment of mass-spectrometry data. To enhance further my computational and statistical skills, and to have the direct access, Professor Benzion Boukai, Ph.D., the chair of the Department of Mathematical Sciences at Indiana University Purdue University Indianapolis agreed to serve as my co-mentor in statistics. I will continue working closely with my mentors and report directly to them.

Research Project

Objectives/Specific Aims:

1. Develop statistical workflows for the analysis of plasma/serum samples obtained from large clinical trials that leverage the specific requirements of clinical cancer researchers, e.g. a more quantitative approach based on the relative quantification of extracted ion chromatograms from mass spectra of control and experimental samples and a restriction of the identification process to differentially expressed peptides and proteins, and the incorporation of times series data.
2. Incorporate additional information on the biological aspects of the data such as biological pathways following from genomics studies, prior information on the studied disease. Test the developed models both on model simulated data and established biological data. Establish sensitivity of the methods as well as their limitations and dependence on the parameters of used technology.
3. Incorporate the developed statistical techniques into the Data Integrative Specimen Information Service including visualization tools/packages to aid clinicians in the decision making.

Studies and Results:

The workflow in development can be applied to any disease. We tested it first on a dataset obtained from plasma samples of patients with cardiovascular disease (CVD). Since CVD is one of the most studied diseases, this dataset presented a great opportunity to validate the results by comparing it with known clinical markers from the literature. The clinical trial for breast cancer is still in the stage of samples' collection, and we plan further to refine our workflow and use it on breast cancer samples. We will test it also on the dataset obtained from a study on mice with breast cancer, and on other available datasets.

We are using patients with cardiovascular disease to test the experimental workflow and computational infrastructure for LC-MS/MS-based proteomic experiments as well as for the NMR and MS based metabolomics approach **(Specific Aim 1, Research Tasks 1-3).** The study cohort involves 3500 patients with coronary artery disease and control patients. Out of this cohort fifty patients, stratified for age and sex, were randomly selected from each of five subgroups; control patients without coronary artery disease, patients with stable angina, patients with unstable angina, patients with non-ST-elevation acute myocardial infarction and patients with ST-elevation acute myocardial infarction. For the LC/MS/MS experiment the high abundance proteins were depleted prior to tryptic digest. Tryptic peptides were analyzed using Thermo-Finnigan linear ion-trap mass spectrometry. Spectra were searched using Sequest and X! Tandem, and signal-processed at Monarch Life Science. All data relevant for further interpretation were indexed and integrated in a single relational database. Overall 2581 proteins were identified and quantified. Of these, 198 proteins were identified with high confidence, 749 with intermediate confidence, and 1634 with low confidence. To generate clinically meaningful results, the protein identifications were mapped to high quality publicly available data sources **(Specific Aim 2, Research Task 4),** including UniProt/SwissProt and NCBI Entrez Gene locus ids and differential quantification was rolled up to the gene level. This improved statistical power and clinical interpretation of the results. The 2581 proteins were combined into, 102 (high confidence identification), 597 (intermediate confidence) and 1317 (low confidence) protein families on the gene level. To achieve a dataset with minimal false identifications, functional annotations from the GeneOntology and the HUPO project were added as well as known plasma concentration from the literature **(Research Task 4)**. Based on this additional information the proteins were assigned a confidence level for the correct identification. Of the 2016 protein families 95 were found to be identified with very high confidence (they were found in the large published HUPO serum protein data set and the serum concentration was in the range that can be measured with LC-MS/MS); 28 were identified with high confidence (12 proteins had borderline reported serum concentrations and 16 proteins were detected in the large HUPO serum protein data set but have not been detected in serum before); 65 protein were identified with intermediate confidence (they had a

known extracellular location but were not present in the HUPO data set and no serum concentration was reported in the literature); and 1828 with low confidence. Figure 1 shows the least abundant proteins identifiable by LC-MS/MS in the plasma, after depletion of high abundance proteins **(Research Task 6)**.

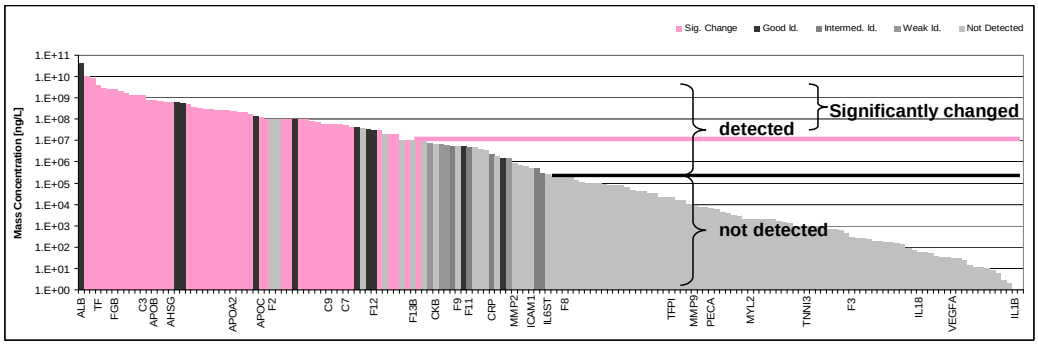


Figure 1: Proteins identified by LC-MS-MS

Peptide abundance was analyzed using the Empirical Bayes approach Limma. Peptide-level quantitative information was combined into protein families using a random-effects model. The list of differentially abundant protein families with a false discovery rate of 0.05 was determined by resampling. The data was annotated with functions from GeneOntology, pathways from Ingenuity Systems, and plasma concentrations from the literature **(Specific Aim 2)**.

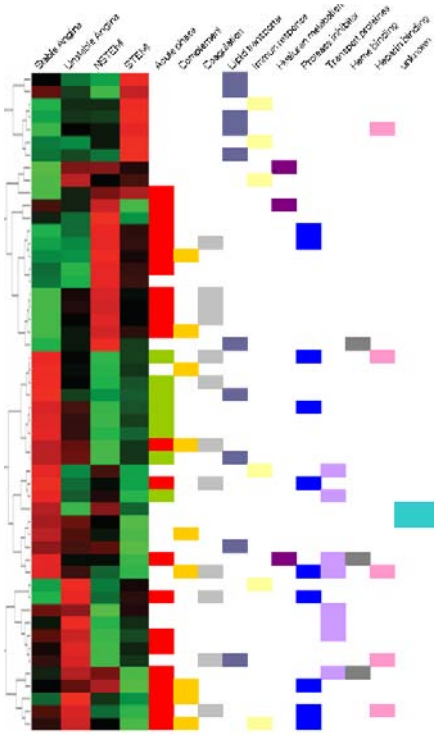


Figure 3 shows an example of clustering across 4 disease stages using LC/MS/MS data

To understand the biological relevance of the differentially expressed proteins we used several functional annotation tools. The commercial network prediction tool Ingenuity Pathway Analysis was used to identify relevant pathways and networks. In addition, all proteins were mapped to networks of protein interactions from IntAct. In the case of protein interaction networks, the sub-networks were determined computationally by means of a graph-theoretical algorithm implemented in Bioconductor. Figure 2 shows an example of a global molecular network developed from information contained in the Ingenuity Pathways Knowledge Base, demonstrating the power of looking at protein networks in the overall context. Figure 3 shows the result of the literature and database searches for the functional annotations and biological processes of the differentially expressed proteins**(Specific Aims 2 and 3, Research Task 4)**. Clustering was done independently of the functional annotation by the pattern of each protein across all groups in Bioconductor using Euclidian distance. Results from this study were validated with nephelometry measurements on 11 proteins **(Research Task 8)**.

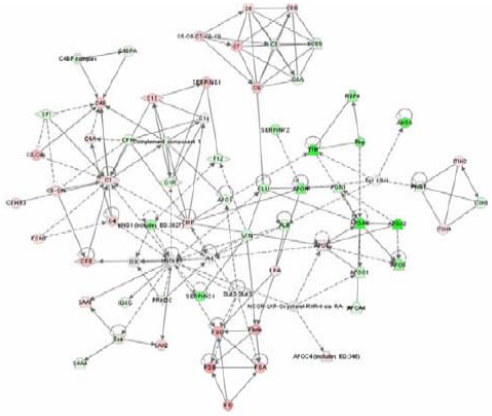


Figure 2: Example of a global molecular network of differentially expressed proteins (red increased; green decreased)

The emerging field of metabolomics promises to add significant new markers to the proteomics approach, and a study of many metabolites together with proteins could become invaluable for a comprehensive pathway and network analysis in cancer. The advantage of analyzing metabolites is that there are much fewer total metabolites in humans (estimated at 3000-5000) compared to proteins. In addition, in response to different physiologic and disease states, metabolites are more prone to perturbations compared to proteins, resulting in changed flux of many metabolites. As part of this project we complemented NMR, which allows for more robust and reproducible measurements, with mass spectrometry-based methods that offer better sensitivity. The NMR experiment resulted in 50 metabolites that were differentially expressed between the 5 groups. The MS data is currently processed and analyzed.

Other Activities

During the last year I mentored several students in the areas related to the project. In order to attract students to this field I have developed and taught the course "Computational Proteomics" in the summer of 2007. The course discussed the modern aspects and tools of high throughput proteomics. To foster the collaboration with researches from industry I have also assisted in the newly organized course BIOS 627 "Statistical Issues in Pharmaceutical Research" taught by researches from Eli Lilly & Company.

Key Research Accomplishments:

- Developed the preliminary version of computational/statistical workflow for analysis of data from clinical studies based on high throughput technologies such as mass-spectrometry and protein microarrays of plasma/blood
- Incorporated prior biological knowledge from biological pathways and public databases
- Tested the developed workflow on patients with cardio vascular diseases
- Compared found proteins with known results from the literature and validated it versus laboratory essay on 11 proteins

Conclusion:

In our study we demonstrate how global proteomic study of plasma can be used to classify the risk of cardiac events. Moreover, we aim the goal of presenting a panel of proteins that can be used for this purpose in the clinical study. Further, to improve the sensitivity we study the biological pathways and plan to include other variables in the study, such as the gene expression data from microarrays and targeted mass spectrometry of proteins of interest and cytokines, as well as clinical data. The developed workflow will be put in use for breast cancer samples when available.

We will complete metabolomic analyses within the next 12 months. We will perform protein microarray analysis on the serum of the 250 patients with cardiovascular disease. We will develop workflows for the analysis of protein microarrays and merge it with mass-spectrometry proteomic analysis. At this point we have available mass-spectrometry data of mice with breast cancer. We will apply our developed proteomic workflow on this dataset. Over the next twelve months we will also continue to improve the Data Integrative Specimen Information Service, and include workflows for the analysis of protein microarrays.

References:

V.Fokin, *"Mass Spectrometry Based Proteomics and Computational Diagnostics: Profiling of Plasma Samples in Coronary Artery Disease",* Open House of The Center for Mathematical Biosciences & The Center for Bio-Computing, October 26, 2007, Indianapolis, IN

Ragg , S., Schadow, G., Vitek, O., Fokin, V., Podgorski, K., Ott, I., Braun, S.L., Kastrati, A.,Schoemig, A. (2007) *Proteomic Profiling of Plasma Samples in Coronary Artery Disease*, American Heart Association Scientific Sessions.

Fokin, V.V. *On Clustering of Proteomic Data*. Abstract in the proceedings of the conference "Intelligent technologies in education, economics and management", Voronezh, Russia, 2007.

Joint meeting of Center for Computational Diagnostics, *"Sample selection for the validation experiment: the role of clinical variables"*, May 1, 2008, Indianapolis, IN

V.Fokin, S.Ragg, G. Shadow, K.Podgorski, O.Vitek, I.Ott, *Computational Workflow Development for the Clinical Application of Proteomic Profiling of Plasma Samples*. Abstract in the proceedings of the Department of Defense Breast Cancer Research Program Conference "Era of Hope", 2008

Appendix 1
Abstracts, and publications:

Ragg , S., Schadow, G., Vitek, O., Fokin, V., Podgorski, K., Ott, I., Braun, S.L., Kastrati, A.,Schoemig, A. (2007) *Proteomic Profiling of Plasma Samples in Coronary Artery Disease*, American Heart Association Scientific Sessions, Abstract 2599:

Introduction: Characterization of the global protein profiles of patients with coronary artery disease has been made possible through advances in protein analytical technologies. Automated label-free relative protein quantification using mass spectrometry (LC/MS/MS) now allows the integration of global protein profiling into large clinical trials.

Methods and Results: In the present study we applied this approach to a subset of samples selected from a consecutive series of 3500 patients that underwent coronary angiography at the Heart Center. Fifty patients were randomly selected from each of five subgroups:

1. ST-elevation acute myocardial infarction,
2. non ST-elevation acute myocardial infarction,
3. unstable angina,
4. stable angina, and
5. control patients without coronary artery disease.

High abundance proteins were depleted prior to tryptic digest. Tryptic peptides were analyzed using Thermo-Finnigan linear ion-trap mass spectrometer. Spectra were searched using Sequest and X! Tandem, and signal-processed at INCAPS. Peptide abundance was analyzed using the Empirical Bayes approach Limma, peptide-level quantitative information was combined into protein families using a random-effects model, and the list of differentially abundant protein families with a false discovery rate of 0.05 was determined by resampling. A total of 1949 protein families were detected and quantified. Of these, 555 protein families had at least one significant change between the groups, including 72 that have been previously associated with cardiovascular disease. The numbers of altered proteins are shown in Table 1.

Conclusion: Our study provides the most comprehensive dataset of protein changes in patients with coronary artery disease described so far. We demonstrate that moderately abundant proteins may potentially be useful for risk classification in coronary artery disease.

Table 1: Protein expression by group compared to normal controls

	STEMI	NSTEMI	Unstable Angina	Stable Angina
Number of proteins with increased levels	88	177	136	90
Number of proteins with decreased levels	46	162	98	76

Fokin, V.V. *On Clustering of Proteomic Data*. Abstract in the proceedings of the conference "Intelligent technologies in education, economics and management", Voronezh, Russia, 2007.

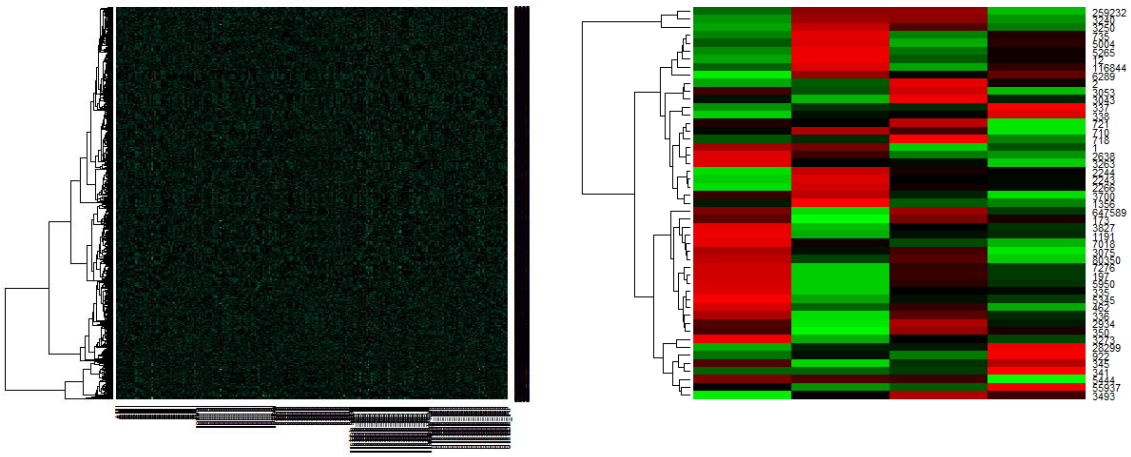
Cluster analysis is one of the commonly used exploratory techniques based on measure of dissimilarity between cases. They are typically grouped into hierarchical and partitioning methods, and usually it is not a straightforward task which technique should be used in a given problem. It is natural to see clustering of results in medical literature, where most researchers would implement one or another type of so-called heatmap and/or dendrogram representation of clustering results. The obvious advantage of these approaches is the visual ability to recognize patterns. However, clustering methods are not always the best visualization technique; moreover, due to the vast variety of clustering methods, in many cases giving different answers, there's a danger of over-interpretation of data, and one should use it only as a hypothesis generating procedure rather than ultimate solution to multivariate problems.

In this presentation we discuss the small piece of our work on how one can use clustering in interpretation of proteomics data. We follow the approach commonly used in the microarray analysis.

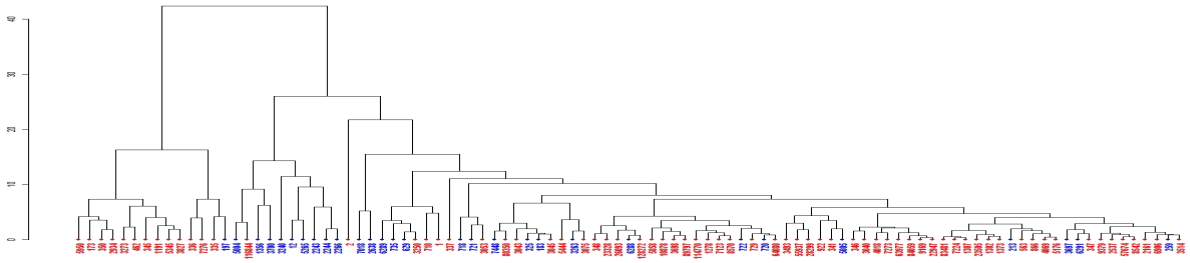
For this study we used the data obtained by the collaborative group of researchers of the Center for Computational Diagnostics at the Indiana University.

Blood was obtained from patients seen at the Heart Center, Munich, Germany over a period of one year. We randomly selected a total of 250 patients for this experiment out of the 3500 who had plasma collected. Fifty patients, stratified for age and sex, were randomly selected from five disease subgroups. Samples were randomized into 2 batches with 5 groups per batch. Randomization was stratified by disease group and age. High Abundance Proteins were depleted from the plasma samples using the commercially available albumin-removal kit (Montage, Millipore). All plasma samples were denatured by 8M urea, digested by trypsin and processed for LC/MS/MS analysis using Thermo-Finnigan linear ion-trap mass spectrometer (LTQ) by the Indiana Center for Applied Protein Sciences.

We have 3 basic layers of data: the raw data (basically the spectra of individual ions from MS), the identified and somewhat quantified peptides, and finally the protein identification and quantification. The raw data, while the vast and extremely complex, is the only "true" dataset; however, in clinical settings one needs to be able to interpret the biological meaning of data such as protein expressions. Moreover, the clustering techniques become obsolete due to the complexity and noise present in the data. The peptide level, while simpler, still doesn't provide either the desired interpretation or ability to use clustering. As a simple example, let us compare two heatmaps below: one done on the peptide level for all patients, and another on selected group of proteins and mean differences in patient groups. The first picture bears no sense at all, while the second can be used for further investigation, for example, to study the possible pathways among the clusters of proteins.



Another approach is to look how different gene (or protein) annotations are distributed among our set of proteins and whether they express similar behavior. For example, for the data in our experiment it is known that acute phase reactants play important role. So we can look if they group together by marking them blue in the dendrogram below.



Obviously some of them cluster together and some do not. So, one can ask whether this information provides any leads toward better understanding of the biology of the disease. In fact, in the case above, one of proteins not marked originally in blue, i.e., not classified as being an acute phase, after the investigation was found to be in the same pathway.

These are only two examples out of many hundreds that typically a researcher would go through while studying similar datasets.

Clustering provides simple way to look for hypothesis. In no way it should be presented as a final answer, and one should be very suspicious of the results claimed on the base of clustering techniques only. In our project we use the tool commonly used in microarrays analysis for the new field of proteomics. While one can not apply it directly, after some modifications it shows great potential.

Abstract in the proceedings of the Department of Defense Breast Cancer Research Program Conference "Era of Hope", 2008

COMPUTATIONAL WORKFLOW DEVELOPMENT FOR THE CLINICAL APPLICATION OF PROTEOMIC PROFILING OF PLASMA SAMPLES

BC060902

Vladimir Fokin,[1] Susanne Ragg,[1] Gunther Schadow,[1] Krzysztof Podgorski,[2] Olga Vitek,[3] and Ilka Ott[4]

[1]Indiana University, Indianapolis, [2]Lund University, Sweden, [3]Purdue University, and [4]Technical University Munich, Germany

Clinical applications of cancer proteomics include disease marker discovery for diagnosis, prognosis, and drug response as well as characterization of signaling and protein pathways. Due to the large dynamic range over which proteins must be detected in plasma (10^{10}), several techniques for identifying differentially expressed proteins have to be combined, such as liquid chromatography followed by mass spectrometry and multiplex antibody arrays. High-throughput technology such as mass spectrometry requires large sample sizes, and the sheer volume of data collected during LC-MS/ MS analysis of plasma samples (1.4 GB) requires that the workflow for data analysis is automated.

One of the major challenges of global proteomics is the lack of rigorous mathematical and statistical tools for analysis of data. While some techniques were developed in genomic studies for microarrays, they cannot be applied directly. For example, in microarrays, one immediately has the intensities for an object of interest, for example, genes, while in the MS/MS proteomics, the quantitative information is obtained for peptides and has to be combined into proteins.

The clinical global proteomics experiment is typically comprised of the following steps:

1. Design of Experiment: A mass spectrometry experiment should be carefully designed to include numerous factors in randomization and stratification scheme.

2. Collection of Samples

3. Preparation of Samples: To increase the sensitivity and to be able to detect the lower abundant proteins, high-abundance proteins are depleted prior to tryptic digestion.

4. LC-MS/MS: The analysis of tryptic peptides is performed by Thermo-Finnigan linear ion-trap mass spectrometer at Monarch Life Sciences.

5. Identification and Quantification of Peptides: The obtained spectra are searched over public databases, and peptides are quantified.

6. Quantification of Proteins: Since the database search does not provide the definite answer for similar proteins, such as in the case of polymorphic protein variances, we combine peptide-level quantitative information into protein families on the gene level.

7. Biostatistical Analysis: Peptide abundances are analyzed using the empirical Bayes approach. The overall significance of multiple hypothesis testing is controlled by the false discovery rate approach. The technique similar to the gene set enrichment analysis in microarrays is used to find statistically significant differences between the protein families. We account for biological dependencies between the proteins by annotating data with functions from GeneOntology and pathways from Ingenuity Systems.

The developed workflow can be applied to any disease. We tested it first on a dataset obtained from plasma samples of patients with cardiovascular disease (CVD). Since CVD is one of the most studied diseases, this dataset presented a great opportunity to validate the results by comparing it with known clinical markers from the literature. The clinical trial for breast cancer is still in the stage of sample' collection, and we plan further to refine our workflow and use it on breast cancer samples.

This work was supported by the U.S. Army Medical Research and Materiel Command under W81XWH-07-1-0447.